

ABSTRACT OF THE DISCLOSURE

The present invention relates to methods and systems for identifying proteins. In particular the invention provides a method for identifying a protein through amino acid sequences of one or more query
5 peptides generated from the protein. The method involves translating amino acid sequences of the query peptides to all possible codons from which the peptides can be synthesized to prepare strings of codons. Known nucleic acid sequences, in particular a set of known nucleic acid sequences including a genome, are searched to locate one or more known nucleic acids that comprise regions that match the strings of codons. Matching nucleic acids are ranked to identify nucleic acids that are true coding regions for the protein to
10 thereby identify the protein.